Japanese Patent Laid-Open (KOKAI) No. H06-223121
Japanese Patent No. 2586372

(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **06223121 A**

(43) Date of publication of application: **12.08.94**

(51) Int. Cl **G06F 15/40**
**G06F 15/40**
**G06K 9/72**

(21) Application number: **05008734**

(22) Date of filing: **22.01.93**

(71) Applicant: **NEC CORP**

(72) Inventor: **KANEDA SATORU**

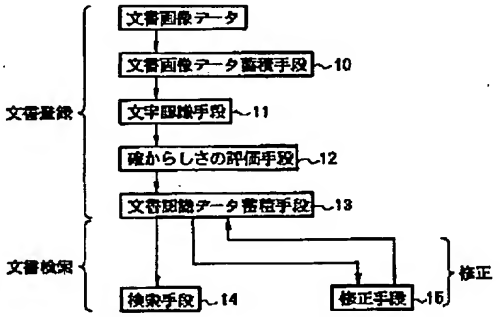
(54) **INFORMATION RETRIEVING DEVICE**

(57) Abstract:

PURPOSE: To prevent retrieval omission caused by uncertain recognition and to perform correction later in the information retrieving device and an information correcting device for performing retrieval and correction with the character recognized result as an object.

CONSTITUTION: A document image is read from a document image data storage means 10, and a character recognizing means 11 divides a character image at every character, performs character recognition and outputs candidate characters and the certainly A evaluating means 12 for certainty selects any candidate so as to prevent retrieval omission from the candidate characters and the certainty, enumerates the candidate and preserves it in list in a document recognition data storage means 13. A retrieving means 14 performs collation for the unit of character on the condition that respective characters are contained in the candidate when reading document recognition data and performing character string retrieval. Therefore, when proper characters are contained in the candidate, the retrieval omission can be prevented even without performing correction.





(19)日本国特許庁（ＪＰ）　(12)特　許　公　報（Ｂ２）　(11)特許番号

第2586372号

(45)発行日　平成９年(1997)２月26日　(24)登録日　平成８年(1996)12月５日

| (51)Int.Cl.[6] | 識別記号 | 庁内整理番号 | Ｆ Ｉ | | 技術表示箇所 |
|---|---|---|---|---|---|
| Ｇ０６Ｆ　17/30 | | | Ｇ０６Ｆ | 15/401 | ３１０Ａ |
| Ｇ０６Ｋ　9/00 | | 9061－5Ｈ | Ｇ０６Ｋ | 9/00 | Ｓ |
| | | | Ｇ０６Ｆ | 15/40 | ３７０Ａ |
| | | | | | ３７０Ｂ |
| | | | | 15/403 | ３１０Ｃ |

請求項の数３(全　８　頁)　最終頁に続く

(21)出願番号　特願平5－8734

(22)出願日　平成５年(1993)１月22日

(65)公開番号　特開平6－223121

(43)公開日　平成６年(1994)８月12日

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　金田　悟
東京都港区芝五丁目７番１号日本電気株式会社内

(74)代理人　弁理士　京本　直樹　(外２名)

審査官　▲吉▼田　耕一

(56)参考文献　特開　平３－50692（ＪＰ，Ａ）
特開　平２－82380（ＪＰ，Ａ）
特開　平３－160580（ＪＰ，Ａ）
特開　昭62－44878（ＪＰ，Ａ）
特開　平４－37971（ＪＰ，Ａ）

(54)【発明の名称】　情報検索装置及び情報検索方法



(57)【特許請求の範囲】

【請求項１】文書画像データを文字認識して得られた文書認識データ群から、任意の検索キーワードを含む文書認識データを検索する情報検索装置において、

入力された文書画像データを蓄積する文書画像データ蓄積手段と、

文書画像データに含まれる文字部分の文字パターンを認識し、候補となる文字コードを選択し、文字コードの確からしさの推定値を求める文字認識手段と、

文字コードの確からしさの推定値の和が、複数の文字コードを格納する文字コードリスト中に正確な文字が含まれるように定めた所定の確率を越えるまで、前記推定値の高い順に候補となる文字コードを前記文字コードリストに追加し、前記文字コードリストの中の候補となる文字コードの数が１つの場合は、先頭の候補となる文字コードを選択し、前記リスト中の文字コードの数が定められたしきい値以内の場合は、候補となる文字コードが複数あることを示す認識コードをともなう先頭の候補となる文字コードを含む複数の候補となる文字コードを選択し、前記リスト中の文字コードの数が定められたしきい値を越えた場合は、候補となる文字コードが多数あることを示す識別コードを選択する確からしさの評価手段と、

これら選択されたコードを文書認識データとして蓄積する文書認識データ蓄積手段と、

入力手段から入力した任意のキーワードを前記文書認識データ蓄積手段から検索する検索手段を備えることを特徴とする情報検索装置。

【請求項２】前記文書認識データ蓄積手段の文書認識データの中の複数の候補となる文字コードから正しい文字

コードの選択や、前記文字コードが多数あることを示す識別コードに正しい文字コードを入力する修正手段を備えることを特徴とする請求項１記載の情報検索装置。

【請求項３】文書画像データを文字認識して得られた文書認識データ群から、任意の検索キーワードを含む文書認識データを検索する情報検索方法において、

入力された文書画像データを蓄積する文書画像データ蓄積ステップと、

文書画像データに含まれる文字部分の文字パターンを認識し、候補となる文字コードを選択し、文字コードの確からしさの推定値を求める文字認識ステップと、

文字コードの確からしさの推定値の和が、複数の文字コードを格納する文字コードリスト中に正確な文字が含まれるように定めた所定の確率を越えるまで、推定値の高い順に候補となる文字コードを前記文字コードリストに追加し、前記文字コードリストの中の候補となる文字コードの数が１つの場合は、先頭の候補となる文字コードを選択し、前記リスト中の文字コードの数が定められたしきい値以内の場合は、候補となる文字コードが複数あることを示す認識コードをともなう先頭の候補となる文字コードを含む複数の候補となる文字コードを選択し、前記リスト中の文字コードの数が定められたしきい値を越えた場合は、候補となる文字コードが多数あることを示す識別コードを選択する確からしさの評価ステップと、

これら選択されたコードを文書認識データとして蓄積する文章認識データ蓄積ステップと、

入力した任意のキーワードを、蓄積された文書認識データから検索する検索ステップを備えることを特徴とする情報検索方法。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は文字データの検索を行う情報検索装置および情報検索方法に関し、特に文書画像から得られた文字データの情報検索装置及び情報検索方法に関する。

【０００２】

【従来の技術】近年、文書情報の増大にともない、文書画像を電子化して蓄積（ファイル）しておき、検索して取り出す装置の需要が高まっている。このような電子ファイル装置において、文字認識を利用して文書画像の文書の内容に基づく検索を可能にする情報検索装置が望まれている。

【０００３】以下、上述した従来の情報検索装置について図を用いて説明する。図５は従来の文字認識を利用した情報検索装置のフローチャートである。

【０００４】文書を登録する際は、文書をスキャナなどの画像入力手段１を利用して文書画像データとして入力し、文書画像蓄積手段２に蓄積する。さらに、文字認識手段３を利用して文書画像データに含まれている文字パ



タン部分を文字認識する。文字認識で認識が不確かな箇所が候補の文字が複数ある箇所については、キーボードなどで登録者が修正作業を行う。この認識結果データを文書画像データと対応づけて認識結果データ蓄積手段６に蓄積する。

【０００５】前記のようにして登録した文書から目的の文書を検索する際は、キーボード７などからキーワードなどの検索条件を入力し、条件を満たす認識結果データを文字検索手段８などにより検索し、認識結果データに対応する文書画像データを出力する。

【０００６】しかしながら、上記のような方式では、文書登録時に上記文字認識後に上記修正作業が必要になり、手間がかかる。

【０００７】また、上記修正作業を行わないと、認識結果データに誤りが含まれる可能性があり、検索対象キーワードを文字列検索するときに、前記キーワードを含む認識結果データに誤りがあると、一致しないため検索漏れが生じる。検索漏れを防ぐために、キーワードの検索時に数文字までの不一致を許す照合手法や、検索文字列と類似したパタン形状の文字列を検索文字列と一緒に検索する手法が考えられてきた。こうした従来例として、信学技報ＣＡ８７－２５（１９８７年5月２９日）、特開平４－１５８４７８号公報等に記載がある。

【０００８】

【発明が解決しようとする課題】しかし、従来技術で行われていた、検索漏れを防ぐために数文字までの不一致を許す照合手法を使うと、不適当な検索結果が生じる（過検索）という問題がある。例えば、１文字までの誤りを許す検索手法で、検索文字列（”自由”など）を検索すると、まったく別の文字列（”理由”や”自然”など）とも一致していると判断してしまう。

【０００９】また、検索文字列と類似形状の文字列とを一緒に検索する手法では、文書画像に書体が異なる文字が含まれている場合など、誤認識の傾向があらかじめ予想されたものと異なる文字が含まれている場合に、検索漏れが起こる。

【００１０】また、従来技術で、検索対象とする文書認識データに認識が不確かであった箇所や認識時の候補が何であったかといった情報が含まれていない場合、このデータだけを用いて修正作業するのは困難である。上記の文書認識データに含まれない情報は、別に保存しておく必要があり、管理が困難である。

【００１１】

【課題を解決するための手段】図１は本発明の構成を示すブロック図である。図１に示すように、上記の課題を解決する第１の情報検索装置は、文書画像データを文字認識して得られた文書認識データ群から、任意の検索キーワードを含む文書認識データを検索する情報検索装置において、入力された文書画像データを蓄積する文書画像データ蓄積手段１０と、文書画像データに含まれる文

字部分の文字パターンを認識し、候補となる文字コードを選択し、文字コードの確からしさの推定値を求める文字認識手段１１と、文字コードの確からしさの推定値の和が、複数の文字コードを格納する文字コードリスト中に正確な文字が含まれるように定めた所定の確率を越えるまで、前記推定値の高い順に候補となる文字コードを前記文字コードリストに追加し、前記文字コードリストの中の候補となる文字コードの数が１つの場合は、先頭の候補となる文字コードを選択し、前記リスト中の文字コードの数が定められたしきい値以内の場合は、候補となる文字コードが複数あることを示す認識コードをともなう先頭の候補となる文字コードを含む複数の候補となる文字コードを選択し、前記リスト中の文字コードの数が定められたしきい値を越えた場合は、候補となる文字コードが多数あることを示す識別コードを選択する確からしさの評価手段１２と、これら選択されたコードを文書認識データとして蓄積する文書認識データ蓄積手段１３と、入力手段から入力した任意のキーワードを前記文書認識データ蓄積手段から検索する検索手段１４を備えることを特徴とする。第２の情報検索装置は第１の情報検索装置に加え、前記文書認識データ蓄積手段の文書認識データの中の複数の候補となる文字コードから正しい文字コードの選択や、前記文字コードが多数あることを示す識別コードに正しい文字コードを入力する修正手段１５を備えることを特徴とする。

【００１２】ここで、確からしさの評価手段１２は、認識が不確かな文字については、候補となる複数の文字コードを列挙して出力する。ただし、候補の数が多い場合は、候補を列挙すると文書認識データが大きくなってしまうため、代わりに、全ての文字が候補であること（候補多数）を示す識別コードだけを出力する。

【００１３】本発明の情報検索方法は、文書画像データを文字認識して得られた文書認識データ群から、任意の検索キーワードを含む文書認識データを検索する情報検索方法において、入力された文書画像データを蓄積する文書画像データ蓄積ステップと、文書画像データに含まれる文字部分の文字パターンを認識し、候補となる文字コードを選択し、文字コードの確からしさの推定値を求める文字認識ステップと、文字コードの確からしさの推定値の和が、複数の文字コードを格納する文字コードリスト中に正確な文字が含まれるように定めた所定の確率を越えるまで、推定値の高い順に候補となる文字コードを前記文字コードリストに追加し、前記文字コードリストの中の候補となる文字コードの数が１つの場合は、先頭の候補となる文字コードを選択し、前記リスト中の文字コードの数が定められたしきい値以内の場合は、候補となる文字コードが複数あることを示す認識コードをともなう先頭の候補となる文字コードを含む複数の候補となる文字コードを選択し、前記リスト中の文字コードの数が定められたしきい値を越えた場合は、候補となる文字コードが多数あることを示す識別コードを選択する確からしさの評価ステップと、これら選択されたコードを文書認識データとして蓄積する文章認識データ蓄積ステップと、入力した任意のキーワードを、蓄積された文書認識データから検索する検索ステップを備えることを特徴とする。

【００１４】

【作用】確からしさの評価手段１２は、文字認識手段１１により文字認識時に得られた文字コードの確からしさの推定値の和が、複数の文字コードを格納する文字コードリスト中に正確な文字が含まれるように定めた所定の確率を越えるまで、推定値の高い順に候補となる文字コードを前記文字コードリストに追加し、前記文字コードリストの中の候補となる文字コードの数が１つの場合は、先頭の候補となる文字コードを選択し、前記リスト中の文字コードの数が定められたしきい値以内の場合は、候補となる文字コードが複数あることを示す認識コードをともなう先頭の候補となる文字コードを含む複数の候補となる文字コードを選択し、前記リスト中の文字コードの数が定められたしきい値を越えた場合は候補となる文字コードが多数あることを示す識別コードを選択する。

【００１５】このため、上記のようにして得られた文書認識データと検索対象キーワードとの文字ごとの照合をする際に、一致条件をキーワードの各文字が文書認識データの候補の中のどれかと一致すればよいとしておくことで、検索漏れを低減することができる。また、画像の条件が良い場合など高い確からしさで認識できる場合は候補を列挙しないので、余分な一致の発生が低減される。

【００１６】また、文字認識の不確かさを評価して、誤りのありそうな箇所とその候補を選んで選択しているため、余分な候補が除去され、文書認識のデータ量が過大になるのが防がれる。

【００１７】また、文書認識データに、認識が不確かな箇所と候補の情報が含まれるため、これを使って後から修正作業を行うことも可能である。

【００１８】

【実施例】以下この発明の実施例について図面を参照しながら説明するが、この発明は以下の実施例に限定されるものではない。

【００１９】図２は本発明の実施例の構成のブロック図を示すものである。図２において文書をスキャナ１６などの画像読み取り手段で、画像データとして読み込み、文書画像データ蓄積手段１７に蓄積する。文字認識手段１８は、文書画像データ蓄積手段１７から画像データを読み出して、文字が書かれた領域を識別して、文字パタンを認識し、候補となる文字コードとその確からしさを推定して確率値で出力する。

【００２０】確からしさを推定するためには、パタンを

処理して得られるいくつかの統計量を軸とするベクトル空間を仮定して、あらかじめいろいろな文字についてその空間での座標を求めておく。そして、文字認識時に認識対象とする文字パタンを、同様に処理して上記ベクトル空間での座標を求め、学習時に近傍にあった座標を見つけ、それらとの距離から推定する。

【００２１】確からしさの評価手段１９は、図３に示すようなアルゴリズムに従う。各文字ごとに、文字認識手段が出力した候補のうち最も有力な候補をリストの先頭に入れる。そして、認識の確からしさをもとに、正しい文字がリストに含まれる確率Ｐを求める。確率Ｐが、あるしきい値Ｐｔｈより小さいならば、候補から次に有力な候補をリストに追加して、確率Ｐを求めなおして繰り返す。選び出されたリストの中の候補の数（リストの長さ）が１個ならば、この文字列コードだけを出力する。リストの中の候補の数があるしきい値の個数Ｎｔｈより少ないなら、候補が複数個あることを示す識別コードと共に出力候補の文字コードを出力する。出力候補の数がＮｔｈを越えていたならば、出力候補の文字コードの代わりに、全ての文字が候補であることを示す識別コード、すなわち、候補多数を示す識別コードを出力する。

【００２２】識別コードの具体的な例としては、正規表現に準拠した記述が考えられる。例えば、候補として｛”Ｂ”、”Ｅ”、”３”｝を列挙する表現は、”［ＢＥ３］”となり、認識不可の表現は、”．”となる。ＮＥＣというパタンを認識した結果、Ｎをはっきり認識し、Ｅの候補が｛”Ｂ”、”Ｅ”、”３”｝で、Ｃが候補多数であった場合、正規表現で記述すると、”Ｎ［ＢＥ３］＊”となる。

【００２３】文書認識データ蓄積手段２０は、評価手段１９が出力する文字コードと識別コードからなるコード列を文字認識データとして蓄積する。

【００２４】文字列検索手段２１は、キーボード２４などの入力手段から入力した検索対象キーワードと文書認識データ蓄積手段２０から読み出した文書認識データとを比較照合し、検索対象キーワードを含む文書認識データを検索する。

【００２５】この文字列照合の例として図４に示したものは、文書認識データのテキストとキーワードを１文字１文字比較して、キーワードの文字列とテキストが全ての文字で一致している場合、キーワードの一致が成立したと判断するものである。ただし、候補複数を示す識別コードがあった場合は、該当するキーワードの文字がここに列挙された候補に含まれてたならば、この文字は一致しているとみなす。また、候補多数を示す識別コードがあった場合は、該当するキーワードの文字が何であっても、この文字は一致しているとみなす。

【００２６】検索結果を知らせるために、上記の検索文字列が含まれる文書認識データ、あるいは、この文書認識データに対応する文書画像データをディスプレイ２２から表示する。

【００２７】文書認識データ修正手段２３は、文書認識データをディスプレイ２２に表示し、複数の候補が列挙されている箇所については、ユーザーにキーボード２４から正しいものを選択される。また、候補が特定されなかった箇所についは、正しい文字コードをキーボード２４から入力させる。

【００２８】

【発明の効果】以上の実施例によれば、第１に、文字パタンの認識時に１つの候補だけでは確からしさを保証できない箇所は、確からしさの推定値が十分になるように複数の文字を候補にしたり、全ての文字を候補として指定するため、認識結果のテキストデータの大きさを過大にすることを防ぐことができる。

【００２９】第２に、認識文字毎の認識の確からしさに応じて候補を列挙するため、検索漏れと過検索の低減を両立することができる。

【００３０】第３に、文書画像に部分的なノイズや異フォントが含まれる場合など、誤認識の傾向が異なるときでも、個別に誤認識の傾向を示すデータ等を用意しなくても検索できる。

【００３１】第４、に修正作業を後から行うことができるため、検索前の修正作業の手間と時間を省くことができる。

【００３２】第５に、これらの修正作業を形態素解析などを行って、候補となる文字を選択した場合、この単語が辞書に含まれるかどうかによって候補を選択するような処理を行う場合も、これらの処理は処理装置に余裕ができるまで後回しできる。

【００３３】第６に、認識結果内に複数の候補を記述する際の仕様が統一されてさえいれば、文字認識手段は異なっていてもかまわないため、英文や手書き文など文書ごとの特性に応じた異なる文字認識手段が作成した認識結果も同一の検索手段で検索することができる。

【００３４】第７に、将来、より高性能な文書認識手段に切り替えた場合でも、それまでに蓄積した文書認識データを継続して利用することができる。

【図面の簡単な説明】

【図１】この発明の構成を示すブロック図である。

【図２】この発明の実施例を示すブロック図である。

【図３】確からしさの評価手段の実施例の動作を示すフローチャートである。

【図４】文字列検索の動作例を示す図である。

【図５】従来技術を示すブロック図である。

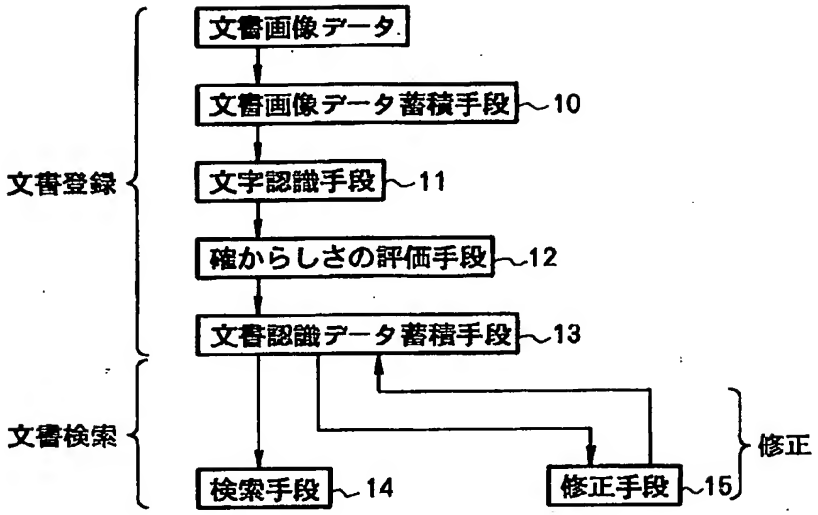
【符号の説明】

１６　スキャナ
１７　文書画像データ蓄積手段
１８　文字認識手段
１９　確からしさの評価手段
２０　文書認識データ蓄積手段

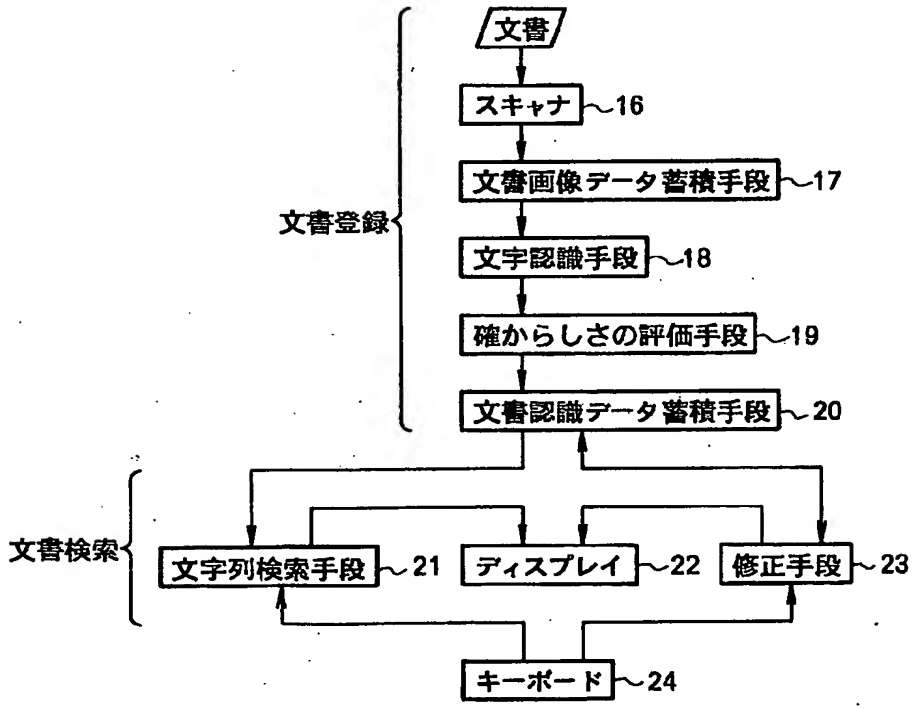
9　　10

2.1　文字列検索手段　　２３　修正手段

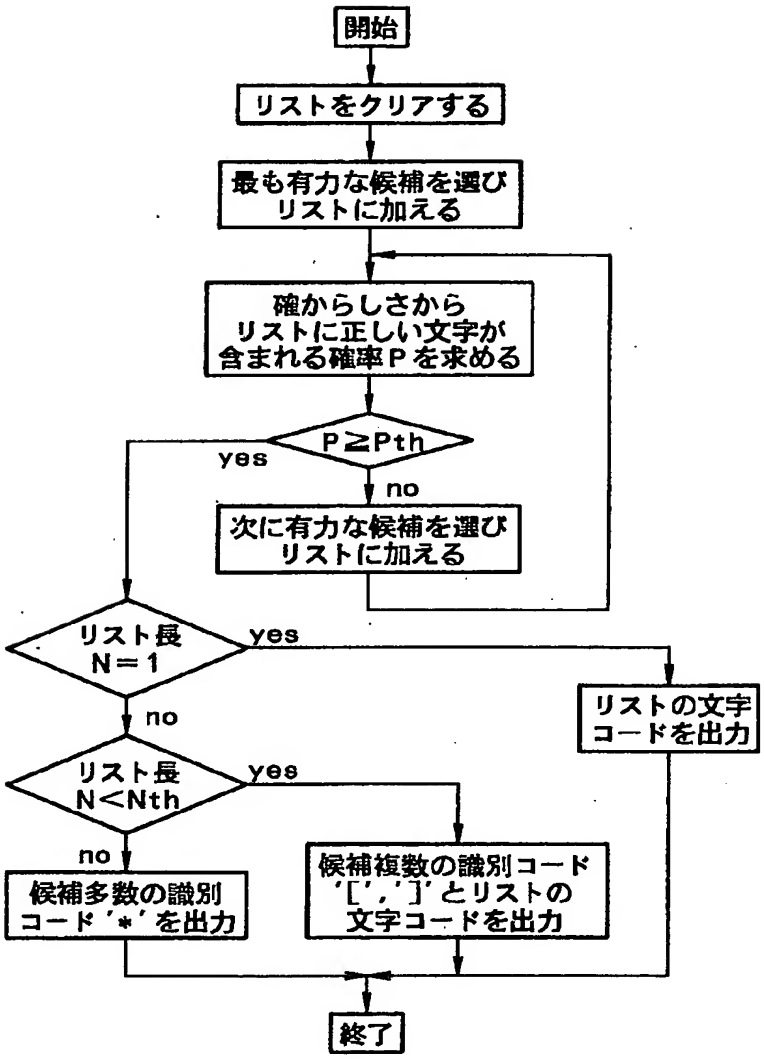
【図１】



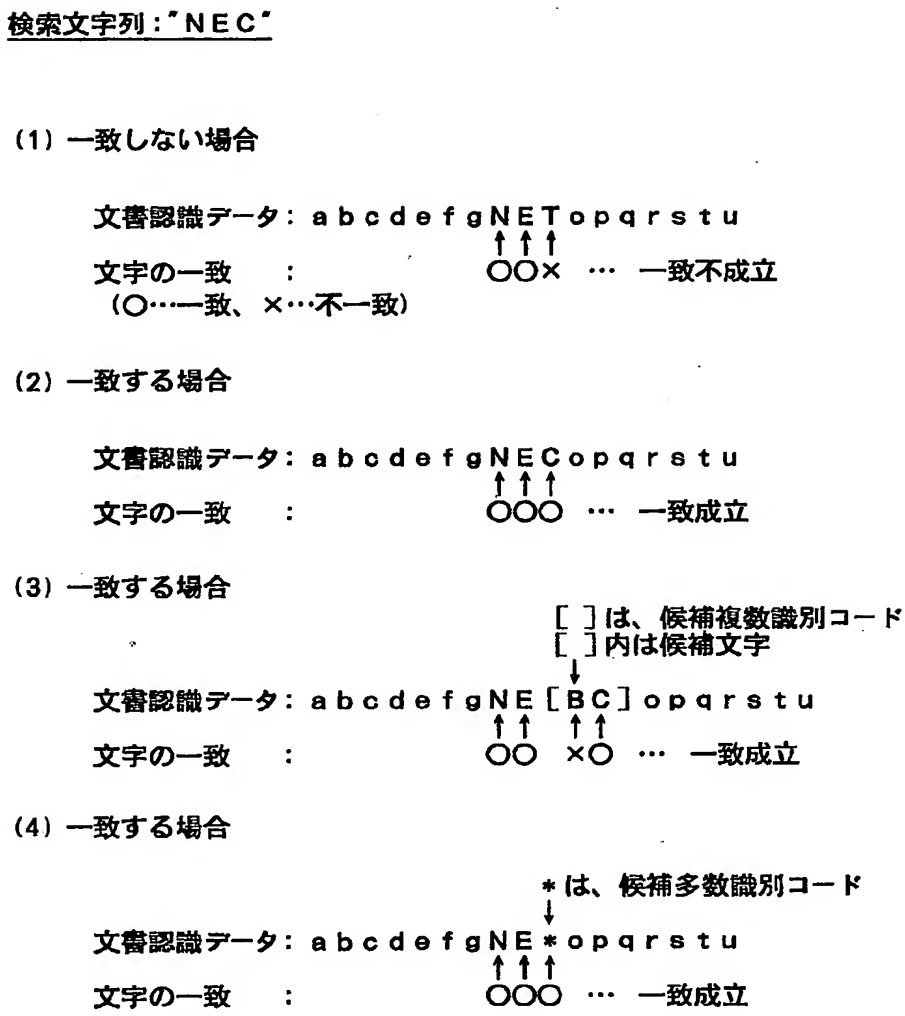

【図２】



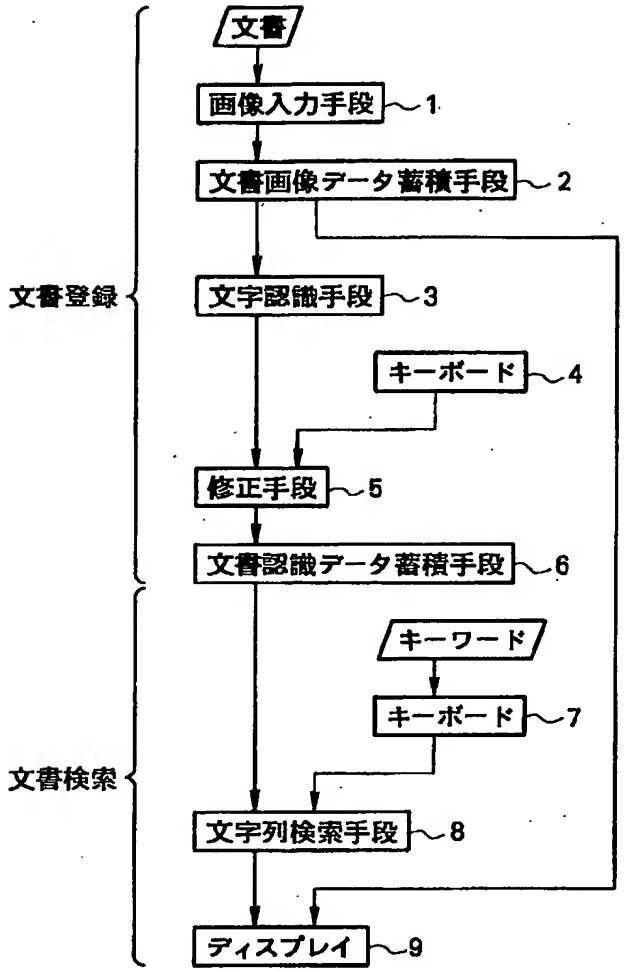
【図３】


開始
リストをクリアする
最も有力な候補を選び
リストに加える
確からしさから
リストに正しい文字が
含まれる確率Ｐを求める
Ｐ≧Ｐth
yes
no
次に有力な候補を選び
リストに加える
リスト長
Ｎ＝１
yes
no
リストの文字
コードを出力
リスト長
Ｎ＜Ｎth
yes
no
候補多数の識別
コード'*'を出力
候補複数の識別コード
'['，']'とリストの
文字コードを出力
終了

【図４】


検索文字列："ＮＥＣ"
(1) 一致しない場合
文書認識データ：ａｂｃｄｅｆｇＮＥＴｏｐｑｒｓｔｕ
文字の一致　：　ＯＯ×　…　一致不成立
（Ｏ…一致、×…不一致）
(2) 一致する場合
文書認識データ：ａｂｃｄｅｆｇＮＥＣｏｐｑｒｓｔｕ
文字の一致　：　ＯＯＯ　…　一致成立
(3) 一致する場合
[ ]は、候補複数識別コード
[ ]内は候補文字
文書認識データ：ａｂｃｄｅｆｇＮＥ［ＢＣ］ｏｐｑｒｓｔｕ
文字の一致　：　ＯＯ　×Ｏ　…　一致成立
(4) 一致する場合
＊は、候補多数識別コード
文書認識データ：ａｂｃｄｅｆｇＮＥ＊ｏｐｑｒｓｔｕ
文字の一致　：　ＯＯＯ　…　一致成立

【図５】




フロントページの続き

| (51)Int.Cl.[6] | 識別記号 | 庁内整理番号 | ＦＩ | 技術表示箇所 |
|---|---|---|---|---|
| | | | Ｇ０６Ｆ　15/403 | ３５０Ｃ |